A History of Western North Pacific Tropical Cyclone Track Forecast Skill

JAMES S. GOERSS AND CHARLES R. SAMPSON

Naval Research Laboratory, Monterey, California

JAMES M. GROSS

Tropical Prediction Center, National Hurricane Center, Miami, Florida

ABSTRACT

The tropical cyclone (TC) track forecasting skill of operational numerical weather prediction (NWP) models and their consensus is examined for the western North Pacific from 1992 to 2002. The TC track forecasting skill of the operational NWP models is steadily improving. For the western North Pacific, the typical 72-h model forecast error has decreased from roughly 600 km to roughly 400 km over the past ten years and is now comparable to the typical 48-h model forecast error of 10 years ago. In this study the performance of consensus aids that are formed whenever the TC track forecasts from at least two models from a specified pool of operational NWP models are available is examined. The 72-h consensus forecast error has decreased from about 550 km to roughly 310 km over the past ten years and is now better than the 48-h consensus forecast error of 10 years ago. For 2002, the 72-h forecast errors for a consensus computed from a specified pool of two, five, seven, and eight models were 357, 342, 329, and 309 km, respectively. The consensus forecast availability is defined as the percent of the time that consensus forecasts were available to the forecaster when he/she was required to make a TC forecast. While the addition of models to the consensus has a modest impact on forecast skill, it has a more marked impact on consensus forecast availability. The forecast availabilities for 72-h consensus forecasts computed from a pool of two, five, seven, and eight models were 84%, 89%, 92%, and 97%, respectively.

1. Introduction

Over the past decade the number of numerical weather prediction (NWP) models capable of producing high-quality tropical cyclone (TC) track forecasts has grown. Today, for the western North Pacific, forecasters at the Joint Typhoon Warning Center (JTWC) routinely use TC track forecasts from eight operational NWP models. Three of these models are run operationally at the Fleet Numerical Meteorology and Oceanography Center (FNMOC): the Navy Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Rosmond 1991; Goerss and Jeffries 1994), the Geophysical Fluid Dynamics Laboratory (GFDL) Hurricane Prediction System (GFDN; Kurihara et al. 1993, 1995, 1998; Rennick 1999), and the Coupled Ocean–Atmosphere Mesoscale Prediction System (COAMPS; Hodur 1997[1]). Two models are run operationally at the Japan Meteorological Agency (Kuma 1996): the global spectral model (GSM) and typhoon model (TYM). The remaining

[1] COAMPS is a trademark of the Naval Research Laboratory.

Corresponding author address: James S. Goerss, NRL, 7 Grace Hopper Ave. Stop 2, Monterey, CA 93943-5502.
E-mail: goerss@nrlmry.navy.mil

three models are the U. K. Meteorological Office global model (UKMO; Cullen 1993; Heming et al. 1995), the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS; Lord 1993), and the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5) run operationally by the Air Force Weather Agency (AFWA; Grell et al. 1995). A time line indicating when TC track forecasts from these eight NWP models became available to the forecasters at JTWC is displayed in Fig. 1.

The benefits of consensus forecasting have long been recognized by the meteorological community (Sanders 1973; Thompson 1977). Leslie and Fraedrich (1990) and Mundell and Rupp (1995) applied this approach to TC track prediction and illustrated the forecast improvement that resulted from using linear combinations of forecasts from various TC track prediction models. Goerss (2000) first illustrated the superior TC track forecasting performance of simple ensemble average or consensus forecasts formed using operational NWP models. For the western North Pacific in 1997 he demonstrated that the TC track forecast errors for both a global-model consensus (GSM, NOGAPS, and UKMO) and a regional-model consensus (GFDN and TYM) were significantly less than the errors for the best of the indi-
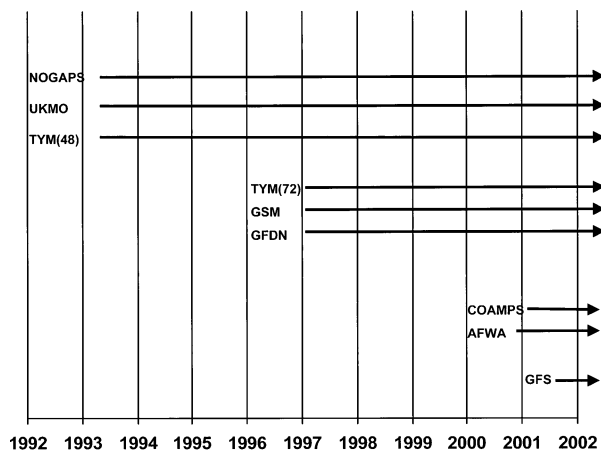
FIG. 1. Time line indicating when TC track forecasts from the operational NWP models became available to the forecasters at JTWC.

vidual models. Versions of these global- and regional-model consensuses were installed on the Automated Tropical Cyclone Forecasting System (ATCF; Sampson and Schrader 2000) in 1998 and run operationally (although somewhat intermittently) during the 1998–2000 seasons. Elsberry and Carr (2000) expanded upon this work and made a five-model consensus (using the models from the Goerss global-model consensus and regional-model consensus) an integral part of the expert system developed for use by the TC forecasters at JTWC, the Systematic Approach Forecasting Aid (SAFA; Peak et al. 2000; Carr et al. 2001). Currently, consensus forecast aids are routinely used by the forecasters at both JTWC and the National Hurricane Center.

In the next section we describe how the TC forecast tracks from the individual NWP models are prepared for use and how a consensus forecast is determined. In section 3, we examine the forecast performance for the western North Pacific of individual NWP models and various consensus forecasts since 1992 and conclude with a summary of our results and a discussion of their implications for the future.

2. Methods

Forecast tracks discussed in this paper are processed as they would be in an operational setting. Since forecast track output for the NWP models become available to the forecaster 6 or 12 h after NWP model run time, they arrive too late to be used directly. Instead, the NWP model tracks are interpolated to intermediate times, and then interpolated positions are relocated to reflect the forecaster-analyzed (best track) position. The version of the interpolator used in this study includes a cubic spline (M. DeMaria 2000, personal communication) and a 10-pass, 3-point filter. All interpolated tracks are computed from real-time tracks, not postseason analyzed tracks (best tracks). Quality control for the interpolator in-

cludes a linear interpolator to fill in missing 12- and 36-h forecasts, a forecast position check (the 6-h/12-h old NWP model 6-h/12-h interpolated forecast position must be within 333 km of the current forecaster analyzed position), and a forecast track speed check (60-kt maximum) for all forecast periods beyond 12 h. NWP model interpolated tracks that fail the 12-h forecast position check are eliminated from the interpolator, while those failing the 60-kt speed check are truncated before the 60-kt speed is encountered.

A consensus for a given forecast period is a simple average of the interpolated members that pass the interpolator quality control tests described above. An attempt is made to compute a consensus forecast at the 12-, 24-, 36-, 48-, 72-, 96-, and 120-h forecast periods. This consensus is computed if *two or more* members exist for a given forecast period. If less than two members exist, the consensus is not computed.

3. Results and conclusions

First, we examine the forecast performance of the interpolated versions of the three NWP models that have been available to the JTWC forecasters since 1992: NOGAPS, UKMO, and TYM. The 12-h forecast errors for the three models from 1992 to 2002 are displayed in Fig. 2a. We see that the forecast errors for the three models declined over the decade from about 120–165 km to just under 100 km. The 24-h forecast errors for the three models declined from about 220–275 km to under 170 km (Fig. 2b). Similar results are seen in Fig. 2c for the 48-h forecast errors, which have improved from about 430–520 km to about 300 km between 1992 and 2002. Finally, we see in Fig. 2d that the 72-h forecast errors declined from roughly 600 km in 1992 to just over 400 km in 2002. Over the decade NWP model forecasts have improved so that 24-h forecasts today are only a little worse than 12-h forecasts in the early 1990s, and 72-h forecasts today are better than 48-h forecasts from the early 1990s.

Forecast performance since 1992 of consensus forecasts created from the interpolated versions of the eight NWP models shown in Fig. 1 was examined. While the total number of models available to these consensus forecasts varies, we only require that forecasts from at least two of the models be available to create a consensus forecast. Henceforth, we will identify the consensus forecasts by the number of models in the specified pool of models from which the consensus forecast is created. For example, as illustrated in Fig. 1, the pool for the three-model consensus consists of NOGAPS, UKMO, and TYM, while the pool for the five-model consensus consists of those three models along with GFDN and GSM. Thus, for the five-model consensus, consensus forecasts can be created when forecasts from two, three, four, or five models are available. For each forecast length, the consensus forecast is merely the arithmetic mean of individual forecasts of available
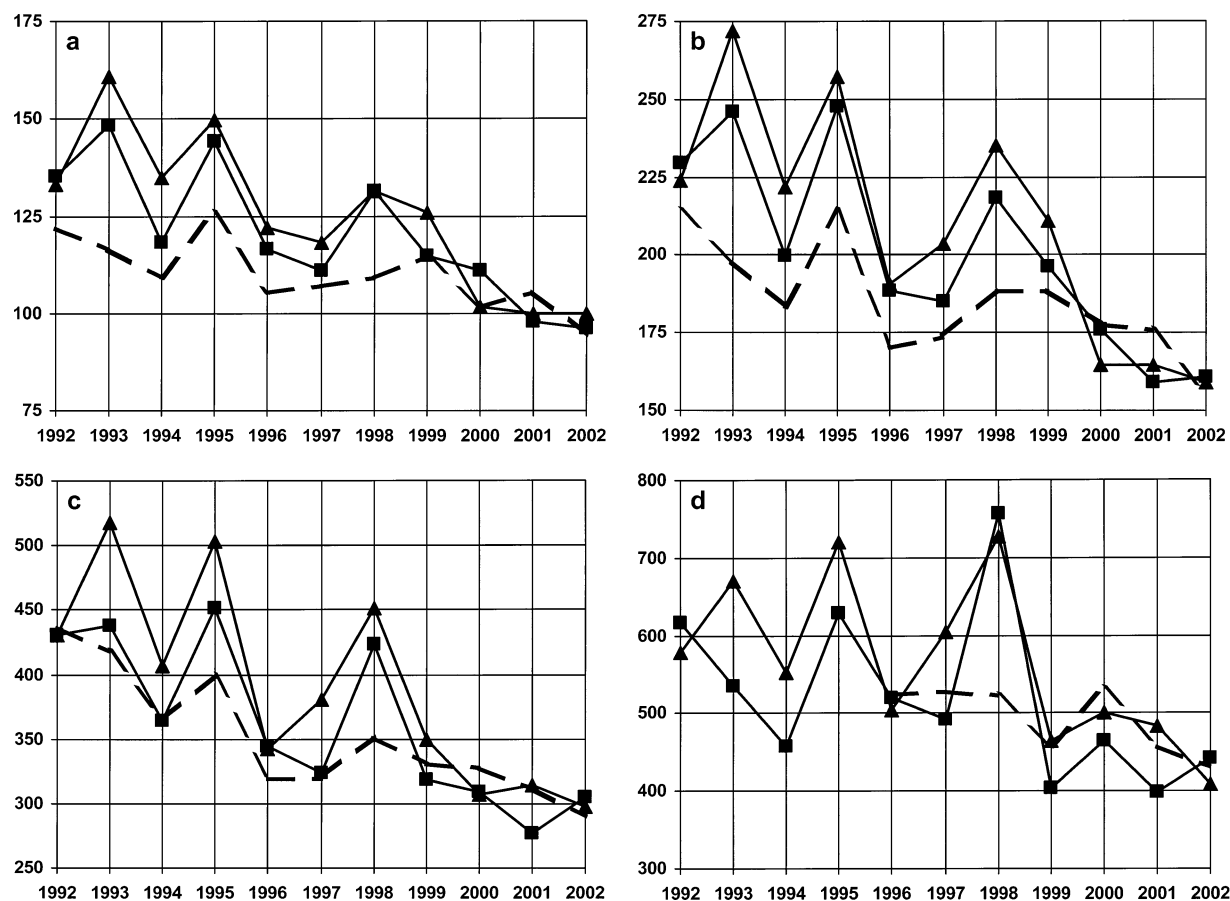
FIG. 2. Mean position errors (km) for NOGAPS (triangles), UKMO (squares), and TYM (dashed line) (a) 12- (b) 24- (c) 48-, and (d) 72-h forecasts for the western North Pacific (1992–2002).

members. The 12-h forecast errors for various consensus forecasts from 1992–2002 are displayed in Fig. 3a. The steady decline in the forecast error of the three-model consensus over the decade from about 120 km to about 80 km is indicative of improvements made to individual NWP models that make up the three-model consensus (NOGAPS, UKMO, and TYM). Making more models available to consensus has resulted in small but consistent gains in skill. For 2002, forecast error for the three-model consensus was about 80 km, while that for the eight-model consensus was just less than 75 km. We see similar results for all forecast lengths. The 24-h forecast error of the three-model consensus declined from about 185 km to about 135 km over the decade (Fig. 3b), while for the eight-model consensus in 2002 it was just less than 120 km. In Fig. 3c, the 48-h forecast error of the three-model consensus declined from about 350 km in 1992 to about 250 km in 2002, while that for the eight-model consensus in 2002 was about 210 km. Finally, the 72-h forecast error of the two-model consensus (NOGAPS and UKMO) declined from about 550 km to about 360 km over the decade (Fig. 3d), while for the eight-model consensus in 2002 it was about 310 km. As we saw for individual NWP models (Fig.

2), over the decade consensus forecasts have improved so that 24-h forecasts today are only a little worse than 12-h forecasts from the early 1990s, and 72-h forecasts today are better than 48-h forecasts from the early 1990s.

We have seen that addition of models to the consensus results in small but consistent gains in skill. For example, in 2002, the errors for the two-, five-, seven-, and eight-model-consensus 72-h forecasts were 357, 342, 329, and 309 km, respectively. However, this is not the only benefit of making more models available to the consensus. By increasing the number of models in the specified pool available to the consensus, we make it more likely that forecasts from at least two models will be available and that a consensus forecast can be formed. In Fig. 4, the availability percentages for the various consensus models in 2002 are displayed. We define the forecast availability percentage to be the percent of the time that consensus forecasts were available to the JTWC forecaster when he/she was required to make a TC forecast. In 2002, the availability percentages for the two-, five-, seven-, and eight-model-consensus 72-h forecasts were 84%, 89%, 92%, and 97%, respectively. By increasing the number of models in the
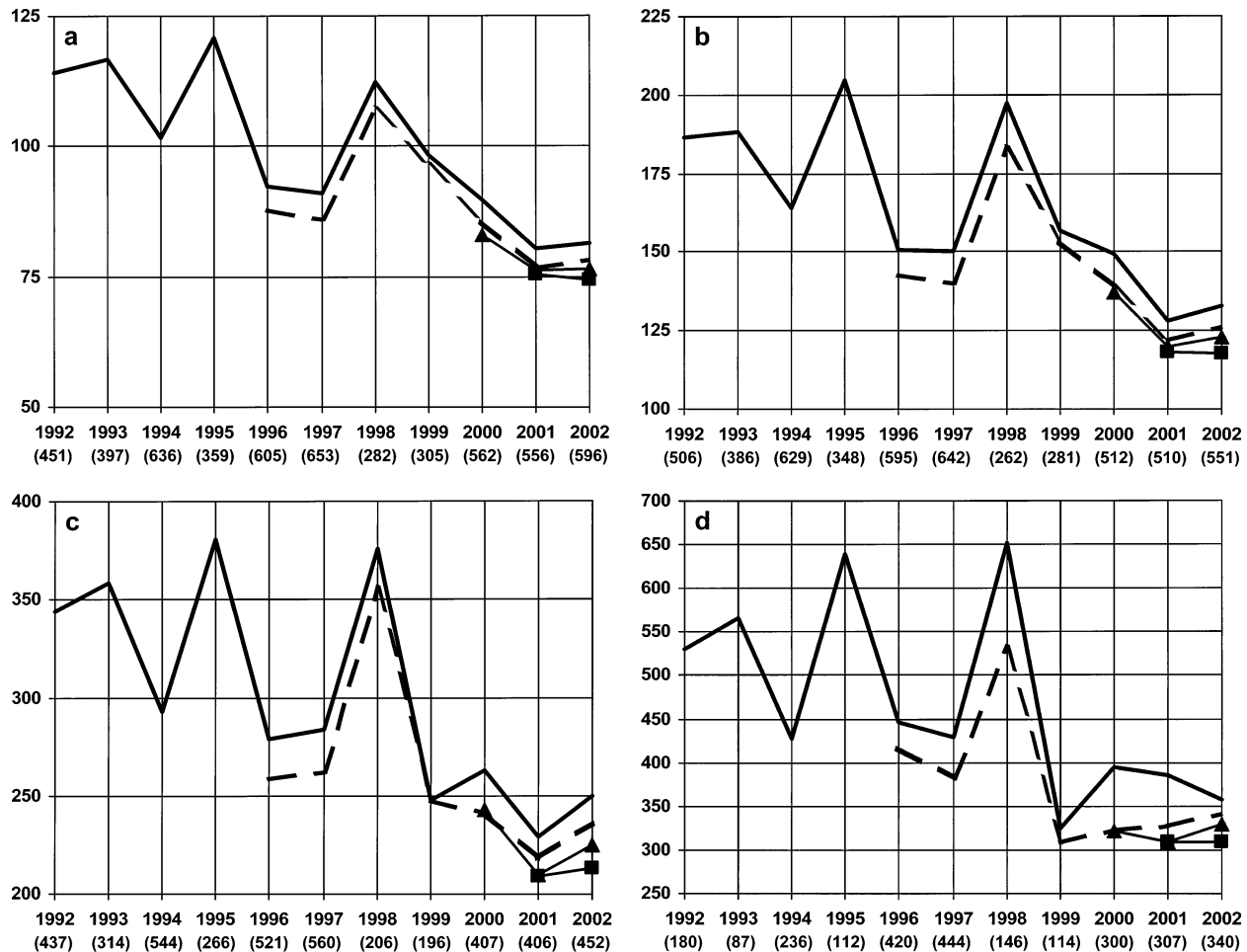
FIG. 3. As in Fig. 2, but for the three-model (solid line), five-model (dashed line), seven-model (triangles), and eight-model (squares) consensuses. The number of cases appears below each year in parentheses.
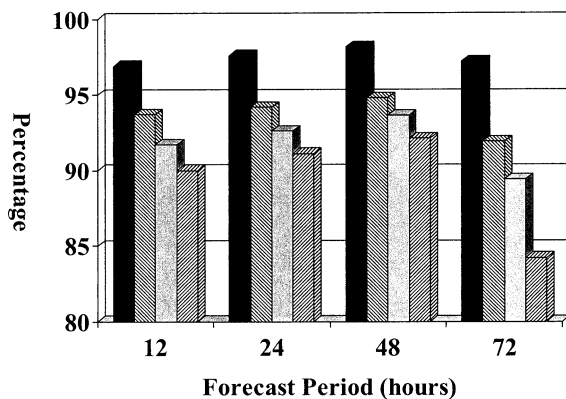

FIG. 4. Forecast availability percentage for the eight-model (solid black), seven-model (descending hatch), five-model (solid gray), and three-model (ascending hatch) consensuses for the western North Pacific in 2002.

specified pool from two to eight, we have significantly increased the percent of the time that forecasts from at least two of the models are available so that a consensus forecast can be created. For all forecast lengths, the availability of the five-model consensus ranged from 89% to 94%, while that for the eight-model consensus ranged from 97% to 98%. To an operational forecaster, this increase in availability may be just as valuable as the increase in forecast skill.

The forecast difficulty level (FDL) for a particular year is defined as the forecast error of a climatology and persistence (CLIPER) model run on best-track initial data (Neumann 1981). The FDLs for the western North Pacific for 1992–2002 were 169, 405, and 611 km at 24, 48, and 72 h, respectively. These values compare quite closely with the values reported by McBride and Holland (1987) of 183, 417, and 632 km. The percent improvement (positive values) or degradation (negative values) with respect to best-track CLIPER of the various consensus forecasts is displayed in Fig. 5. For the three-model consensus the 12-h forecast skill in-
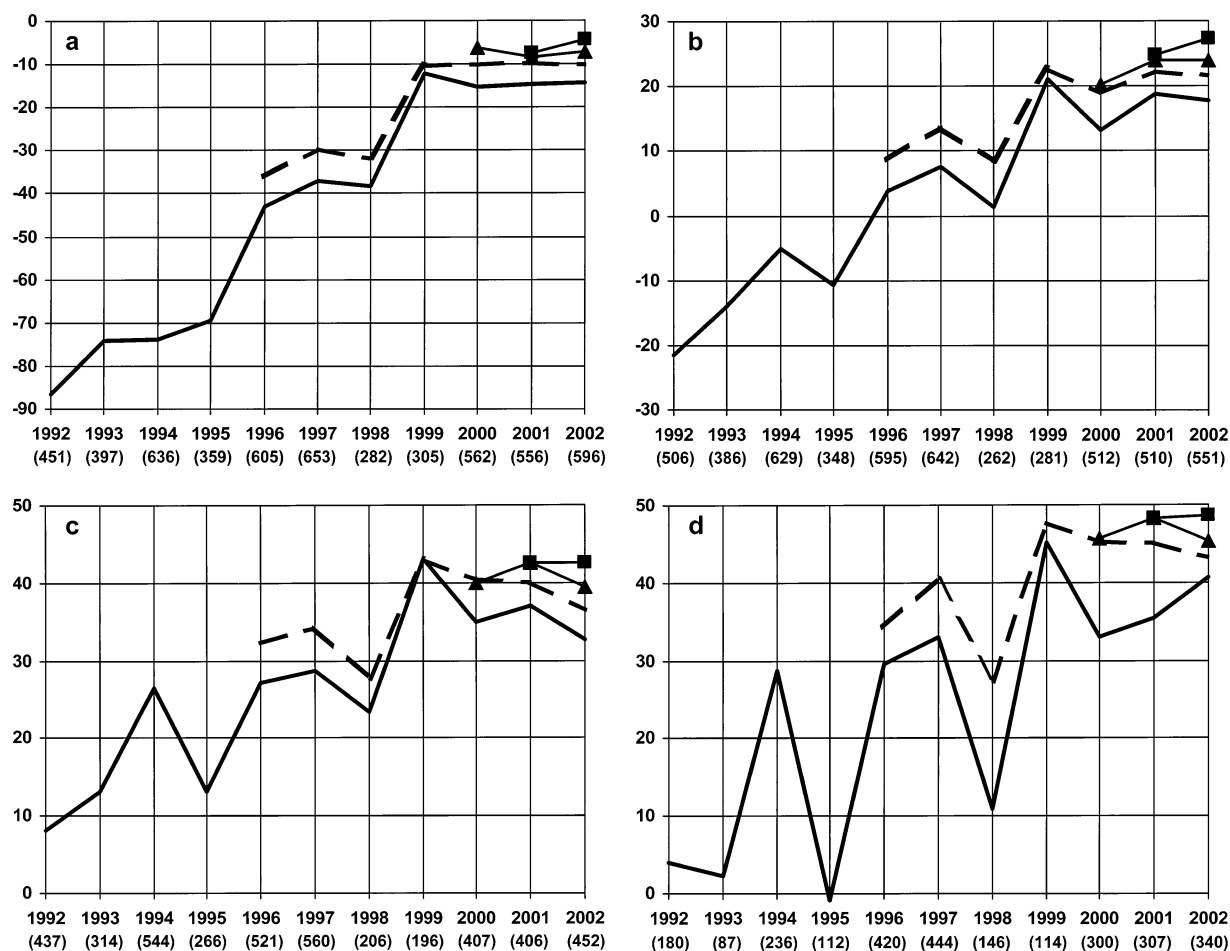
FIG. 5. As in Fig. 3, but for percent improvement (indicated by a positive number) or degradation (indicated by a negative number) with respect to best-track CLIPER. The number of cases appears below each year in parentheses.

creased from nearly −90% in 1992 to about −15% in 2002 (Fig. 5a). Once again, we see that making more models available to the consensus has resulted in small but consistent gains in skill. For 2002, 12-h forecast skill increased from about −15% for the three-model consensus to about −5% for the eight-model consensus. The 24-h forecast skill for the three-model consensus increased from about −20% in 1992 to nearly 20% in 2002 (Fig. 5b), while the forecast skill for the eight-model consensus in 2002 was greater than 25%. In Fig. 5c, we see that the 48-h forecast skill for the three-model consensus increased from less than 10% to greater than 30% over the decade, while that for the eight-model consensus in 2002 increased further, to greater than 40%. Finally, the 72-h forecast skill for the two-model consensus increased from less than 5% in 1992 to about 40% in 2002 (Fig. 5d), while that for the eight-model consensus increased further, to nearly 50%.

In conclusion, we have seen that the TC track forecasting skill of NWP models for the western North Pacific has improved dramatically over the past decade (1992–2002). This improvement has contributed to a

similar improvement in consensus forecasts created from these NWP models. We have also seen that the addition of models to the consensus results in improvements to both consensus track forecast skill and consensus forecast availability. Finally, we have seen the forecasters successfully integrate consensus forecasting into their operational procedures at JTWC. While a consensus demonstrates superior performance with respect to the individual NWP models in the long run, forecasts from individual models can certainly outperform a consensus in the short run. The forecasters use consensus as a baseline or starting point, then modify the forecast as they see fit. In the future, as the TC track forecasting skill of NWP models continues to improve and more high-quality NWP models become available, we can look forward to further improvement in consensus track forecast skill and forecast availability.

vortex trackers, and the entire staff at JTWC for advice and the care and feeding required to make a consensus work in an operational setting. This work was supported by the Oceanographer of the Navy through the program office at the Space and Naval Warfare Systems Command (PMW-150), Program Element 0603207N.

REFERENCES

Carr, L. E., R. L. Elsberry, and J. E. Peak, 2001: Beta test of the systematic approach expert system prototype as a tropical cyclone track forecasting aid. *Wea. Forecasting,* **16,** 355–368.

Cullen, M. J. P., 1993: The Unified Forecast/Climate Model. *Meteor. Mag.,* **122,** 81–122.

Elsberry, R. L., and L. E. Carr, 2000: Consensus of dynamical tropical cyclone track forecasts—Errors versus spread. *Mon. Wea. Rev.,* **128,** 4131–4138.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.,* **128,** 1187–1193.

——, and R. A. Jeffries, 1994: Assimilation of synthetic tropical cyclone observations into the Navy Operational Global Atmospheric Prediction System. *Wea. Forecasting,* **9,** 557–576.

Grell, G. A., J. Dudhia, and D. R. Stauffer, 1995: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398+STR, 122 pp.

Heming, J. T., J. C. L. Chan, and A. M. Radford, 1995: A new scheme for the initialization of tropical cyclones in the UK Meteorological Office global model. *Meteor. Appl.,* **2,** 171–184.

Hodur, R. M., 1997: The Naval Research Laboratory's Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Mon. Wea. Rev.,* **125,** 1414–1430.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the Navy Operational Global Atmospheric Prediction System's spectral forecast model. *Mon. Wea. Rev.,* **119,** 1786–1815.

Kuma, K., 1996: NWP activities at Japan Meteorological Agency. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J15–J16.

Kurihara, Y., M. A. Bender, and R. J. Ross, 1993: An initialization scheme of hurricane models by vortex specification. *Mon. Wea. Rev.,* **121,** 2030–2045.

——, ——, R. E. Tuleya, and R. J. Ross, 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.,* **123,** 2791–2801.

——, R. E. Tuleya, and M. A. Bender, 1998: The GFDL hurricane prediction system and its performance in the 1995 hurricane season. *Mon. Wea. Rev.,* **126,** 1306–1322.

Leslie, L. M., and K. Fraedrich, 1990: Reduction of tropical cyclone position errors using an optimal combination of independent forecasts. *Wea. Forecasting,* **5,** 158–161.

Lord, S. J., 1993: Recent developments in tropical cyclone track forecasting with the NMC global analysis and forecast system. Preprints, *20th Conf. on Hurricanes and Tropical Meteorology,* San Antonio, TX, Amer. Meteor. Soc., 290–291.

McBride, J., and G. Holland, 1987: Tropical-cyclone forecasting: A worldwide summary of techniques and verification statistics. *Bull. Amer. Meteor. Soc.,* **68,** 1230–1238.

Mundell, D. B., and J. A. Rupp, 1995: Hybrid forecast aids at the Joint Typhoon Warning Center: Applications and results. Preprints, *21st Conf. on Hurricanes and Tropical Meteorology,* Miami, FL, Amer. Meteor. Soc., 216–218.

Neumann, C., 1981: Trend in forecasting the tracks of Atlantic tropical cyclones. *Bull. Amer. Meteor. Soc.,* **62,** 1473–1485.

Peak, J. E., L. E. Carr, and R. L. Elsberry, 2000: Developing a systematic approach to tropical cyclone track forecasting expert system—1: Information management, visualization, and proactivity considerations. Preprints, *24th Conf. on Hurricanes and Tropical Meteorology,* Ft. Lauderdale, FL, Amer. Meteor. Soc., 502–503.

Rennick, M. A., 1999: Performance of the Navy's tropical cyclone prediction model in the western North Pacific basin during 1996. *Wea. Forecasting,* **14,** 3–14.

Sampson, C. R., and A. J. Schrader, 2000: The Automated Tropical Cyclone Forecasting System (Version 3.2). *Bull. Amer. Meteor. Soc.,* **81,** 1131–1240.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.,* **54,** 1171–1179.

Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.,* **105,** 228–229.